Choosing priors in bayesian measurement invariance modeling: A Monte Carlo Simulation study

Pokropek, Artur ; Schmidt, Peter ; Davidov, Eldad

Abstract: Multi-group Bayesian structural equation modeling (MG-BSEM) gained considerable attention among substantive researchers investigating cross-group differences and methodologists exploring challenges in measurement invariance testing. MG-BSEM allows for greater flexibility by applying elastic rather than strict equality constraints on item parameters across groups. This, however, requires a specification of user-defined prior variances for cross-group differences in item parameters. Although prior selection in general Bayesian settings is well-studied, guidelines with respect to tuning the normal prior variances in MG-BSEM approximate measurement invariance (AMI) analysis are still largely missing. In a Monte Carlo simulation study we find that correctly specifying prior variances results in more precise credibility intervals (CI) and posterior standard deviations, while prior misspecification has little influence on point estimates. We compared the BIC, DIC, and PPP fit measures and found in our simulation scenarios that the DIC measure was most effective, when a proper threshold for model selection was applied.

DOI: https://doi.org/10.1080/10705511.2019.1703708

Choosing Priors in Bayesian Measurement Invariance Modeling: A Monte Carlo Simulation Study

Artur Pokropek

Institute of Philosophy and Sociology of the Polish Academy of Sciences

Peter Schmidt

Department of Political Science and Centre for Environment and Development (ZEU), Justus Liebig University Giessen

Eldad Davidov

Institute of Sociology and Social Psychology, University of Cologne;

Department of Sociology and University Research Priority Program Social Networks, University of Zurich

Correspondence should be sent to Dr. Artur Pokropek, Institute of Philosophy and Sociology, Polish Academy of Sciences, Nowy Świat 72, Warsaw, 00-330. Phone: 0048501764306; Email: artur.pokropek@gmail.com

Acknowledgments: This work has been prepared under the project Scales Comparability in Large Scale Cross-Country Surveys, which is funded by the Polish National Science Centre, as part of the grant competition Sonata 8 (UMO-2014/15/D/HS6/04934). Eldad Davidov would like to thank the University of Zurich Research Priority Program Social Networks. The authors would like to thank Lisa Trierweiler for the English proof of the manuscript.

Abstract

Multi-group Bayesian structural equation modeling (MG-BSEM) gained considerable attention among substantive researchers investigating cross-group differences and methodologists exploring challenges in measurement invariance testing. MG-BSEM allows for greater flexibility by applying elastic rather than strict equality constraints on item parameters across groups. This, however, requires a specification of user-defined prior variances for cross-group differences in item parameters. Although prior selection in general Bayesian settings is well-studied, guidelines with respect to tuning the normal prior variances in MG-BSEM approximate measurement invariance (AMI) analysis are still largely missing. In this article, we examine how different prior specifications affect the results of MG-BSEM analysis across several conditions. In a Monte Carlo simulation study we find that correctly specifying prior variances results in more precise credibility intervals (CI) and posterior standard deviations, while prior misspecification has little influence on point estimates. We compared the BIC, DIC, and PPP fit measures and found in our simulation scenarios that the DIC measure was most effective to determine whether appropriate priors were selected, when a proper threshold for model selection was applied. Finally, we examined the difference threshold for BIC, DIC, and PPP that informed when to stop increasing the prior in our different scenarios.

Keywords: measurement invariance (MI), approximate measurement invariance (AMI), multi-group confirmatory factor analysis (MG-CFA), multi-group Bayesian structural equation modeling (MG-BSEM), cross-group comparisons, Monte Carlo simulation study, BIC, DIC, PPP, tuning priors, Mplus

Choosing Priors in Bayesian Measurement Invariance Modeling: A Monte Carlo Simulation Study

In recent years a new approach in measurement invariance testing emerged (Muthén & Asparouhov, 2012; Muthén & Asparouhov, 2013). This approach claims that the dichotomy of exact (full or partial) invariance vs. non-invariance in multi-group analyses could be supplemented by the concept of *approximate* measurement invariance (AMI): This allows small "harmless" differences in factor loadings and item intercepts across different groups that do not bias substantive conclusions when performing multiple-group modeling. In this methodology, non-invariance between item parameters from different groups is treated as a ubiquitous and inevitable consequence of between-group differences that could be incorporated into the statistical modeling using Bayesian methods and the so-called multi-group Bayesian structural equation modeling (MG-BSEM) approach (Muthén & Asparouhov, 2013). In these models, "wiggle room" (van de Schoot et al., 2013) between item parameters is introduced by hyperparameters: priors that define the level of invariance between item parameters across groups. Hyperparameters are "elastic" equality constraints that relax the assumption of the full invariance model under demanding conditions of applied social research (Braeken & Blömeke, 2016; Lek et al., 2019; Seddig & Leitgoeb, 2018). These conditions encompass often relatively large sample sizes and a high number of groups (e.g., 25 to 100), particularly when analyzing one or more time points in international cross-country surveys. Under such conditions it is rather unlikely to reach full exact measurement invariance, so allowing measurement parameters to vary a little across groups is both realistic and encouraging for applied researchers who seek data that are sufficiently invariant across their groups and allows meaningful comparisons.

This strategy gained considerable attention in the research community, stimulating high expectations among applied researchers working with real data, methodologists, and statisticians (Braeken & Blömeke, 2016; Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014; Kim, Cao, Wang, & Nguyen, 2017; Pokropek, Davidov, & Schmidt, 2019; van de Schoot et al., 2013; Zercher, Schmidt, Cieciuch, & Davidov, 2015). Although some methodological work has been done to evaluate the performance of the model fit of MG-BSEM models (Muthén & Asparouhov, 2012; Pokropek et al., 2019; van de Schoot et al., 2013), and although considerable work has been published on how to improve the selection of priors in general applications using Bayesian statistics (van de Schoot et al., 2011; van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017; van Erp, Mulder, & Oberski, 2018; Zondervan-Zwijnenburg, Peeters, Depaoli, & van de Schoot, 2017), guidelines for applied researchers with respect to the selection of the priors in MG-BSEM AMI testing are still largely missing (Seddig & Leitgoeb, 2018). The selected priors in the AMI testing should allow the model to provide accurate point and interval estimates of group parameters of interest (e.g., latent means) and their distributions.[1]

Choosing Prior Cariances in MG-SEM AMI Modeling

Muthén and Asparouhov (2012) proposed and further elaborated (Asparouhov, Muthén, & Morin, 2015) a strategy that recommends starting with very small prior variances (e.g., 0.001) when testing for AMI, and then increasing the prior variance multiple times consecutively. This strategy is grounded on a "subjective selection" based on monitoring two criteria:

1. speed of convergence (number of iterations) and

[1] There is a large body of literature on prior selection (see, e.g., Berger & Sun, 2008). In this paper we only refer to prior selection in AMI. We avoid the term "*optimal* prior choice". Instead, we prefer the terms "tuned prior variance" of the group differences in MG-CFA models or simply "the correct or appropriate prior choice" for several reasons. First, our study covers a selected set of conditions, but there are many other possible conditions and scenarios not examined in our study. So, our guidelines may not apply to other conditions. Second, our guidelines, as will be shown later, are different for different conditions. Third, our study is based on Monte Carlo simulations, and thus does not provide any general theoretical proof about the "optimality" of the chosen prior.

2. ninety-five percent confidence interval for the difference between the observed and the replicated chi-square values.

Despite the fact that it is conceptually a good heuristic, its implementation requires extensive experience in working with MG-BSEM estimation, and under certain conditions it may even somewhat resemble an art form or educated guessing more than a grounded statistical procedure. Other authors suggested to focus more on comparative fit indices like the Bayesian information criterion (BIC) and the deviance information criterion (DIC) (Stromeyer, Miller, Sriramachandramurthy, & DeMartino, 2015) or other fit measures like Bayesian root ean square error of approximation (BRMSEA) (Hoofs, van de Schoot, Jansen, & Kant, 2018) when deciding which priors to choose. Using such fit measures is, however, not straightforward because of the special nature of the MG-BSEM parameters. Small-variance prior parameters are not actual parameters like in the frequentist approach, because they are indeed *approximate equality* constraints across groups of measurement parameters (e.g., Muthén & Asparouhov, 2012; Hoijtink & van de Schoot, 2017). Therefore, although controversial (Gelman & Shalizi, 2013), it has been argued that one should apply a different procedure in MG-BSEM models with informative priors that penalizes for complexity (Gelman, Meng, & Stern, 1996; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). Stromeyer et al. (2015) suggested giving more weight to the BIC than to the Akaike information criterion (AIC) when evaluating MG-BSEM models with different prior settings, while Asparouhov et al. (2015) find this recommendation misguided, showing that BIC can unnecessarily penalize the model by counting small-variance prior parameters as actual parameters. Asparouhov et al. (2015) argue that the DIC provides, in fact, a more accurate decision guidance, because the model complexity penalty of the DIC is based on the number of estimated parameters in MG-BSEM models.

Because guidance in the statistical literature on how to evaluate the fit of MG-BSEM models with different priors is still often not clear enough, in real-life applications many researchers choose priors rather arbitrarily, for example, following specifications provided in articles introducing those methods in best-case scenarios. In these articles it is suggested to perform robustness checks with different prior values, in an attempt to determine how sensitive model parameters of interest (usually latent mean rankings) are with different priors (Braeken & Blömeke, 2016; Cieciuch, Davidov, Algesheimer, & Schmidt, 2018). This procedure might not be the best strategy because it may lead to the selection of models with sub-optimal fit and a mismatch between the prior and the true variance, as we demonstrate below.

Goals of the Present Study

In this article, we examine how one may choose priors in *MG-BSEM AMI modeling* by treating priors as a form of regularization (van Erp, Oberski, & Mulder, 2019) rather than as information on cross-group differences between measurement parameters based on previous knowledge. Steck and Jaakkola (2003) explain regularization in the Bayesian approach as specifying a prior distribution over the parameters to subsequently guide the selection of model structures. Using Monte-Carlo simulation studies, we demonstrate the effect of prior misspecifications on selected fit parameters of MG-BSEM. We evaluate different strategies of prior variance selection (i.e., tuning the prior variance of the measurement parameter group differences), pointing to the effective ones under conditions close to real-life situations of applied researchers who use survey data.

The goals of this paper are thus twofold: (1) In Study 1 we illustrate how different prior specifications affect the results of MG-BSEM analysis; and (2) in Study 2 we examine the effectiveness of procedures that aims to tune the normal prior variance of the group differences

in MG-CFA models using model selection criteria. More precisely, Study 2 examines the difference threshold for BIC, DIC, and PPP that inform when to stop increasing the prior variance in AMI testing in our different scenarios, so that a realistic description of cross-group differences is reached, allowing researchers to recover the true parameters of interest.[2]

MG-BSEM Measurement Invariance Analysis

CFA and its multi-group extension (MG-CFA) allow estimating group-specific parameters, where group means are of particular interest for applied researchers. They define the relation between an observed continuous indicator Y_{ig} and the latent trait η_{jg} as a linear equation (for a simple one-dimensional case):

$$y_{ig} = \tau_{ig} + \lambda_{ig}\eta_{jg} + \varepsilon_{ig} \quad (1)$$

where τ_{ig} is the intercept and λ_{ig} the factor loading of an item i in group g. The index j denotes the person, and ε_{ig} is a random error. The intercepts are the expected mean value of Y_{ig} when $\eta_{jg} = 0$.

When all item parameters (factor loadings and intercepts) are set equal across groups and the model is supported by the data, it implies that exact scalar invariance is achieved and mean estimates of latent factors are comparable (e.g., Meredith, 1993; Steenkamp & Baumgartner, 1998). When only the factor loadings are equal across groups (metric invariance), only the covariances and unstandardized regression coefficients can be compared. However, exact (full or partial) *scalar* equivalence is rarely achieved with real survey data (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014). MG-BSEM relaxes assumptions about exact invariance of

[2] This guidance is obviously limited given the finite and small number of conditions included in our study.

the item parameters thus allowing for small cross-group discrepancies (or "wiggle room") in item parameters (Muthén & Asparouhov, 2013; van de Schoot et al., 2013). In other words, whereas item intercepts and loadings are fixed to be *equal* across all groups ($\tau_{ig} = \tau_{ig'}$ and $\lambda_{ig} = \lambda_{ig'}$) in MG-CFA models, in the approximate invariance approach, applying MG-BSEM models, these constraints are relaxed based on the assumption that item-related parameters are *approximately* equal ($\tau_{ig} \approx \tau_{ig'}$ and $\lambda_{ig} \approx \lambda_{ig'}$).

Typically, in MG-BSEM models, non-informative priors are used for all parameters except for the parameters defined for the allowed wiggle room in item measurement parameters (this is the current default when using Mplus 7.2 or later versions [Muthén & Muthén, 1998-2019] for the test). In practical terms, while most parameters are freely estimated, the size of the expected items' measurement parameter differences must be predefined by the user using a prior. In the implementation in Mplus, the differences between item parameters are expressed in terms of a standard normal distribution, with a mean of zero and a difference variance that needs to be predefined, usually in the range of 0.001 and 0.1.

Technically speaking, the difference variance is defined in Mplus by the covariance between two prior distributions of the item parameters. For example, let us take two item intercepts: τ_{11} – the intercept of item 1 in group 1 and τ_{12} – the intercept of item 1 in group 2. In MG-BSEM models, item parameters receive uninformative priors with a zero mean and a large variance: $Prior(\tau_{11}) \sim N(0,1000)$; $Prior(\tau_{12}) \sim N(0,1000)$. Elastic constraints (i.e., the difference variance priors) are imposed by setting a high covariance between two prior distributions, for instance: COV[$Prior(\tau_{11}), Prior(\tau_{12})$] = 999.995. As the variance of the difference between two distributions is defined by the sum of the variances of the two distributions minus two times the covariance ($V(a-b) = V(a) + V(b) - 2COV(a,b)$, where v

is the variance and *a* and *b* are the parameters), the prior variance for the differences between item parameters becomes 1000 + 1000 - 2*999.995 = 0.01.

Those priors aim to reflect small cross-group differences across item parameters in real data. If differences between parameters are the same between all groups, a prior variance of 0.01 in an MG-BSEM model implies a cross-group parameter variance of 0.005, a prior variance of 0.05 implies a cross-group parameter variance of 0.025, etc.[3] Figure 1 presents expected *cross-group* variations of item parameters for different prior variances.

<<< Figure 1 around here >>>

The gray vertical line represents, in the Bayesian approach, a prior variance cross-group difference of 0. This corresponds with the exact classical (frequentist) MG-CFA model to test for measurement invariance, where item parameters are assumed to be exactly the same across groups (i.e., with zero differences, see, e.g., Muthén & Asparouhov, 2013). A prior variance of zero in the MG-BSEM model will thus reduce, in practice, the MG-BSEM model into an MG-CFA model that uses Bayesian estimation techniques rather than maximum likelihood (ML). Prior variances in the Bayesian approach that are higher than zero allow the incorporation of cross-group parameter differences into the model. For instance, a small prior variance of 0.005 indicates that 95% of the cross-group differences would be bounded by -0.14 and 0.14, a prior variance of 0.05 indicates that 95% of the cross-group differences would be bounded by the -0.44 and 0.44 interval, and a prior variance of 0.1 indicates that 95% of the cross-group differences would be bounded by the -0.62 and 0.62 interval (Muthén & Asparouhov, 2012). In fact, higher prior variances like 0.05 or 0.1 (in standardized metric) go beyond the definition of "small cross-group discrepancies" and imply rather high levels of differences between item

[3] In contrast to prior variances, item parameters in different groups are assumed to be independent of each other (with a zero covariance between them), therefore, $V(a-b) = V(a) + V(b)$, where *v* is the variance and *a* and *b* are the parameters.

parameters.

The task of choosing priors might also be described as defining a model that seeks the golden middle between model fit and imposing cross-group measurement equality constraints that allow meaningful comparisons of parameters of interest while still reflecting, to some extent, the true cross-group measurement parameter differences, as depicted in Figure 2. In this figure we illustrate a simple scenario with two groups and five measurement item parameters, for example, item intercepts. The horizontal axis represents the values of the parameters in group 1. The vertical axis represents the values of the parameters in group 2.

<<< Figure 2 around here >>>

The scenario depicted in the first panel (a) reflects an exact invariance model: The parameters in the two groups are constrained to be exactly equal and lie exactly on a straight line. This model maximizes invariance but, in most situations, sacrifices model fit. The last panel (d) depicts a configural invariance model where no equality constraints on measurement parameters are imposed (there is no line defining parameter equality), and parameters are estimated independently for the two groups. This model maximizes the fit to the data but does not ensure comparability of the parameters of interest. The two panels in the middle of Figure 2 describe elastic constraints with different levels of elasticity defined by different values of difference priors. Here, we do not have straight lines but rather regions that define the possible item parameter locations. The region on the second panel is much closer to the straight line than the region on the third panel, because the variance of the prior differences in the second panel is lower than in the third panel. The item parameters in the second panel are closer to the line than those represented in the third panel. Choosing the prior implies, thus, defining such a model that balances the model fit with the invariance requirements.

The estimation of MG-BSEM models relies on Bayesian procedures where parameters are treated as random components. In the estimation procedure, the likelihood function of the data is combined with the prior distribution for each parameter and final estimates are obtained as posterior distributions of the parameters, conditional on the data (for details of the estimation algorithms see, e.g., Gelman et al., 2013; Kruschke, 2015). In general, Markov chain Monte Carlo (MCMC) algorithms are commonly used to obtain posterior distributions. This technique makes random draws of parameter values conditional on some sets of other parameters generating a large number of draws. It then uses them to empirically estimate the posterior distributions. The algorithm applied in the software package Mplus (Muthén & Muthén, 1998-2018) which we use in this paper is an MCMC algorithm based on the Gibbs sampler (see Gelman et al., 2013). In the Gibbs approach, the posterior distributions are obtained by iteratively sampling from the conditional densities of each set of model components with the remaining variables fixed to their current values (for a detailed discussion, see Lee, 2007; for the implementation in Mplus, see Asparouhov & Muthén, 2010).

As mentioned earlier, the possibility to use zero variance priors and diffuse or non-informative priors (rather than informative, small variance priors) allows the Bayesian procedure to produce estimates that resemble those of exact invariance or configural invariance models, respectively, when using ML or least squares estimations (Gelman et al., 2013). The advantage of Bayesian estimation is, however, that it does not require distributional assumptions, and it is easily applicable even to very complex models with many parameters. Hiwever, this is accompanied by the higher price of extensive computational burden and usually a long estimation time (Kruschke, 2015). On the other hand, priors might be informative and incorporate prior beliefs or information beyond the data that is used for the analysis. Indeed, the

use of specific priors in AMI modeling can be justified by two major arguments. First, measurement theories in the social sciences have not reached a precision level that would allow postulating exact invariance across groups. Thus, small uncertainty with respect to the similarity or dissimilarity of measurement parameters across groups should be taken into account by priors. Second, it was demonstrated by past simulations that, within certain limits, small variance priors may not bias the estimation of parameters of interest such as latent means, covariances, or regression coefficients (e.g., van de Schoot et al., 2013). MG-BSEM models combine the use of diffuse and informative priors. Whereas for most parameters diffuse priors are used, prior beliefs about the level of invariance of the item parameters (e.g., factor loadings and intercepts) are introduced by small variance informative priors.

As Bayesian estimation is based on different principles than the frequentist estimation approach, common fit indices (West, Taylor, & Wu 2012), such as the chi-square goodness-of-fit statistic, the RMSEA, or the comparative fit index (CFI), to evaluate the model fit are not available. In the Bayesian framework it is, however, possible to formulate comparative fit indices like AIC and BIC to evaluate the fit of Bayesian models. These comparative fit indices are also known and often used for evaluating the fit of frequentist, non-Bayesian models (West et al., 2012),

$$AIC = -2\log\left[p\left(y\middle|\hat{\theta}\right)\right] + 2k$$

$$BIC = -2\log\left[p\left(y\middle|\hat{\theta}\right)\right] + k\log(n)$$

where y is the observed data, $\hat{\theta}$ is the likelihood estimate of the parameters, k is the number of parameters in the model, and n is the sample size. Those indices are expected to work well if non-informative or zero variance priors are used. However, when using informative priors, establishing the number of parameters is problematic (Spiegelhalter et al., 2002). AIC and

BIC can unnecessarily penalize the MG-BSEM model by counting small-variance prior parameters as actual parameters and thereby overshadowing or obscuring information provided by the model. As an alternative measure, the DIC index was developed for the Bayesian approach (Spiegelhalter et al., 2002). It is formulated similarly to the AIC, with one difference in the formula: The estimated number of parameters k_D replaces the number of parameters k. Technically, the estimated number of parameters is the posterior mean of the deviance minus the deviance of the posterior means (see Gelman et al. 2013, pp. 172-173). In linear models with non-informative priors, AIC and DIC are expected to be equal:

$$DIC = -2\log\left[p\left(y\middle|\hat{\theta}\right)\right] + k_D$$

The estimated number of parameters does not regard small-variance prior parameters as actual parameters, and therefore DIC outperforms AIC and BIC measures, at least in MG-BSEM analysis (Asparouhov et al., 2015).

Another way to assess model fit in a Bayesian framework is to use the posterior predictive p-value (PPP; Gelman et al., 1996). In the PPP, the chi-square for the observed and replicated (or updated) data is subsequently computed for each iteration within the Markov chain (in the Mplus implementation for every 10th iteration). In MG-BSEM, PPP is the proportion of iterations for which the replicated chi-square exceeds the observed chi-square. A “good” fit is achieved if the PPP is around 0.50. Values under 0.50 indicate an underfit of the model whereas values larger than 0.5 indicate an overfit of the model. The PPP was found to be robust for assessing model fit within small samples, but also sensitive due to trivial deviations from the hypothesized model with large samples (Muthén & Asparouhov, 2012). Although some authors have criticized the PPP measure (Hoijtink & Van de Schoot, 2017), the validity of its use as a fit measure has not been convincingly challenged (for a discussion, see Asparouhov & Muthén

2017).

Although there exists some research on model fit indexes (see, e.g., Spiegelhalter et al., 2002), to the best of our knowledge, the performance of BIC, DIC, and PPP for the evaluation of AMI, that is, for evaluating whether appropriate priors were chosen, has not been examined in-depth. In this study we will explore their performance under different conditions as outlined in the following section.

Monte Carlo Simulations

Three scenarios reflecting three common (or alternative) types of research design are examined in this study (see Table 1).

<<< TABLE 1 around here >>>

In each scenario we examined six levels of (true) variances in measurement parameter differences (factor loadings and intercepts simultaneously): 0.000 (i.e., exact invariance), 0.001, 0.005, 0.010, 0.025, and 0.050 together with six levels of prior specifications. Each true level of variance was confronted with six levels of priors (i.e., with one correct match and five mismatches) constructing 36 conditions. In the first scenario we used 4,000 replications per condition, resulting in 144,000 (4,000 x 36) estimations of MG-BSEM. In the remaining scenarios with larger number of groups and sample sizes we used 400 replications per condition, resulting in 14,400 estimations of MG-BSEM. Additionally, for each condition, the classical MG-CFA scalar model was estimated to compare it with the MG-BSEM models. This model was used as a baseline for the comparisons as it fully constrained the parameters to be equal across groups thus ignoring the problem of measurement invariance. Thus, approximate invariant

models that considered parameter differences across groups were expected to perform better than this model in the presence of non-invariance. In each scenario we used a single latent variable measured by five items for which the reliability, as measured by Cronbach's alpha, was 0.9.[4]

Data Generating Procedures

Data for the simulations were generated using a CFA model for continuous data. In the first step, we sampled means and standard deviations for each group from normal distributions N(0,0.3) and N(1,0.1), respectively. We chose these distributions after examining distributions and cross-country differences of latent means obtained from MG-CFA modeling of real data from scales of political trust, openness to experience, social engagement, and attitudes toward immigrants from the European Social Survey (see www.europeansocialsurvey.org).

In the second step, we generated parameters for each item. Factor loadings were sampled from a uniform distribution bounded by a mean of 0.6 to 0.7. Those values are considered rather high in survey research (Brown, 2015, p. 130, e.g., recommended that factor loadings should be at least as high as 0.3 or 0.4 to be considered reliable). However, these loadings guarantee that the majority of the produced factor loadings in the simulation would lie between 0.3 and 1 after allowing them the wiggle room to differ across groups. Intercepts were drawn from a standard normal distribution N(0,0.5) that mimics a situation with high variability across item intercepts.

In the third step, using sampled distributions and item parameters, we generated continuous data that fulfilled the assumption of exact measurement invariance.

In the fourth step, we added approximate non-invariance bias to the item parameters. The bias was added to measurement parameters (factor loadings and intercepts) using random draws from a standard normal distribution with a mean of zero and a variance that depended on the

[4] Although Cronbach's alpha has been criticized in the literature (see, e.g., Brown, 2015; Sijtsma, 2009), we use it here because it is still much in use by applied researchers.

simulation conditions. For instance, for simulating AMI at the level of 0.05 (i.e., a situation where the distribution of the differences between parameters has a mean of zero and a variance of 0.05), the bias for each item was drawn from a standard normal distribution with a mean of 0 and a variance of 0.025 (i.e., the variance of the differences of two random variables equals the sum of their variances assuming a covariance of zero between them).

In the final step, data were generated using parameters obtained from step 4, reflecting a situation where AMI is present.

For all computations, we used the Mplus 7.2 computer program (Muthén & Muthén, 1998-2018). For MG-CFA models, we used the maximum likelihood estimation with robust standard errors (MLR) with the default Mplus estimation criteria (Muthén & Muthén, 1998-2018). The Gibbs sampling with two MCMC chains was used. A process was assumed to converge when the second half of the iterations had potential scale reduction (PSR) convergence criteria values lower than 0.01 (for technical details, see Asparouhov & Muthén, 2010). A minimum number of iterations was set to 5,000, and a maximum number of iterations was set to 100,000.

The convergence rate for all the models under all conditions with a prior variance smaller than 0.05 was virtually 100% whereas for a prior variance of 0.05, the convergence rate was close to 98%.

Measures of Model Recovery

In our simulation study, we first wanted to determine how different prior specifications affect the ability of the model to recover the true group latent means and standard deviations and to provide consistent rankings of the group latent means and accurate interval estimations. To examine the ability of the model to recover the true parameters and provide consistent rankings, we used six criteria:

1) the correlation of the true means with the estimated means as an indicator of a correct recovery of the group latent mean ranking;
2) the root mean square error (RMSE) to assess the overall accuracy of group means recovery;
3) the mean absolute bias of group means;
4) empirical standard errors of group means;
5) the RMSE to assess the overall accuracy of group standard deviation recovery; and
6) the coverage of the true means by 95% of the mean estimation credibility intervals (CI) generated using posterior standard deviations (SD) of the estimated means in Bayesian estimation, or the coverage based on classical coefficient intervals in maximum likelihood (ML) estimation.

With respect to the first criterion, according to recommendations by Muthén and Asparouhov (2013, 2014), a correlation between the true means and their estimates of at least 0.98 (preferably 0.99) indicates a reasonably good recovery of the group mean rankings. When reporting such correlations, we refer below to *mean correlations*. A good model would provide both a high mean correlation (i.e., a reasonably good recovery of the group mean ranking) as well as a reasonable precision of the point estimates of the latent means.

With respect to the second and fifth criterion, the overall accuracy of the parameter estimates (group means and group standard deviations) is measured by the RMSE. The interpretation of the RMSE is straightforward: the smaller its value, the better is the parameter recovery.

The third criterion examines the mean absolute bias of group means. The bias is defined as the difference between an estimated group mean and the true value over the replications. The

bias reports to what extent the mean estimate is under- or overestimated with respect to true value. We report the mean absolute bias over all groups. The smaller the value is, the lower is its average.

The fourth criterion is the precision of estimates defined as the empirical standard error of the estimates. The magnitude of precision is only dependent on the estimated values and is independent of the true value. It could be interpreted as the statistical variance of an estimation procedure. The higher the reported empirical standard error, the lower the precision of estimates is.

Finally, with respect to the sixth criterion, we assess the coverage of the true means by the 95% coefficient intervals when applying models based on ML estimation, and the 95% credibility intervals generated using posterior standard deviations of the estimated means when applying models with a Bayesian estimation.

The six measures listed above are commonly used indicators in simulation studies (for a comprehensive overview and formal definitions, see, e.g., Boomsma, 2013; Morris, White, & Crowther, 2019; Walther & Moore, 2005).

Study 1: The Effect of Different Prior Specifications on the Results of MG-BSEM Analysis

Simulation Scenario 1

Figure 3 presents sets of matrices in the form of heatmaps for the first scenario. The first panel (a) presents the mean correlations, the second (b) displays the RMSE for means, panel (c) presents the mean absolute bias for the means, panel (d) depicts the empirical SE, panel (e) reports the 95% coverage for the latent means, and panel (f) displays the RMSE for the standard deviations. Dark shades represent more desirable results in terms of our criteria (higher correlations of mean rankings, lower RMSEs, a lower mean absolute bias, lower empirical SE of

means, and more precise 95% CIs) and light shades represent less desirable results. In the columns, we display the different true levels of variance (AMI) in the data we generated. Rows indicate whether a simple MG-CFA model is applied (in the first row) or the applied priors for the MG-BSEM models (in the following rows). On the diagonal of the matrix (excluding the first CFA row) we have situations where priors were specified appropriately, that is, in accordance with the empirical variation in the parameters, reflecting the true level of AMI (i.e., a good match). Off diagonal elements represent misspecifications: The lower left part contains results with priors higher than the actual AMI, and the upper right part contains priors that underestimate the actual level of AMI. For instance, in panel (a), the value of 0.961 in the last row of the first column reports the correlation between the true means and the estimated latent means when the level of AMI was 0.000 but priors for the item differences were set to 0.050 (i.e., the priors were higher than the actual level of AMI).

<<< Figure 3 around here >>>

The results demonstrate the ability of the models to recover the true latent means and standard deviations, the true latent mean rankings, and the means' coverage in each condition. First, it becomes evident that the recovery of parameters is not very precise and does not allow for a reliable recovery of group rankings as explained below. The top left table in the figure (panel a) suggests that using different priors does not result in different correlations of the mean rankings, as evidenced in the similar correlations in the different rows within each column. The point estimates of group means and standard deviations are relatively independent of the chosen priors under various levels of true variances, as evidenced in panels a, b, and f in Figure 3 by the rather similar correlations and RMSE values within each column for the different priors (rows).

This suggests that robustness checks of prior specifications based on correlating point estimates under different priors has little value for identifying the correct priors. Representing mean absolute bias, panel c in Figure 3 shows a different story. The least biased results are generated by either the lowest or the highest priors. Overall, however, the level of bias is very small, and these differences should not be overstated. Indicating the precision of estimates, the empirical SE in panel d also show that the precision generally decreases once the true level of AMI increases. In contrast to the above findings, the bottom-left panel (e) of Figure 3 does demonstrates that the 95% CI of group parameters is sensitive to the prior specification. Thus, the real importance of a "correct" prior specification is allowing a correct estimation of CIs which in turn would enable researchers to make correct statistical inferences.

Simulation Scenario 2

In Figure 4 we present the results for scenario 2 in exactly the same way as for scenario 1. In contrast to the first scenario, in the second scenario, with larger sample sizes, the recovery of latent means allows also for a reasonably accurate mean rankings recovery, when priors are well tuned and the level of AMI is rather low (smaller than 0.005). The effect of incorrect prior specifications is not very large but nevertheless noticeable in the two scenarios. Similarly to scenario 1, the recovery of group point estimates is independent of the chosen priors under various levels of true cross-group variances. Interestingly, according to panel (c) it seems that specifying high prior variances (0.05) in scenario 2 might substantially bias the results, if the priors do not meet empirical reality. According to panels (d) and (e), prior specifications might affect the posterior standard deviation of the estimates in the AMI model and, consequently, bias the interval estimation. Well-tuned priors allow for a more accurate coverage of the mean

estimation as displayed in panel (e).

<<< Figure 4 around here >>>

Simulation Scenario 3

Finally, the results in scenario 3 presented in Figure 5 replicate the results obtained for scenarios 1 and 2 in most aspects. The recovery of means measured by mean correlations (panel a) and the overall accuracy of mean estimates measured by the RMSE (panel b) depends mostly on the true level of AMI and could not be compensated by well-tuned priors. The bias seems to be rather small in all instances, and there is no clear pattern here (panel c). The last three panels (d, e, f) suggest that researchers can potentially gain more accurate estimates in MG-BSEM in situations where priors are well tuned.

<<< Figure 5 around here >>>

Study 2: Tuning the Normal Prior Variance

The analyses so far have shown that choosing incorrect priors may have consequences, particularly in terms of the coverage of the true mean estimates. Next, we will examine how to choose priors using the global fit measures, discussed in the previous section, in our specific scenarios. We specifically examine how efficient the three fit measures (BIC, DIC, and PPP) are in recognizing whether prior variances should be further tuned.

Muthén and Asparouhov (2012) suggested that in the search for an appropriate prior, one should begin with the simplest model (with a prior that equals zero) and then gradually increase its prior until a significant improvement of the model fit is achieved. However, they did not define which improvements of fit should be considered as significant. In the context of model comparison in Bayesian analysis, Cain and Zhang (2018) suggested that $\Delta PPP > 0.10$ or 0.15 and $\Delta DIC > 7$, 5, or 3 imply a considerable change in model fit depending on the sample size and

complexity of the model. However, they did not examine MG-BSEM models that explre AMI with different priors nor did they examine models with larger sample sizes as is commonly the case in large-scale cross-country studies, which make it hard to generalize their conclusions.

Using data from our simulation studies and the same conditions in the three scenarios, we utilized a stepwise model selection strategy starting from a model with a zero prior, and we compared this model with a model with a higher prior (i.e., 0.001). If the fit (BIC, DIC, PPP) of the model with the higher prior (i.e., 0.001) was better, we continued and compared it to a model with an even higher prior (i.e., 0.005) and so on. A model was chosen when the improvement of fit in terms of BIC, DIC, and PPP did not exceed a certain threshold. Our goal was to assess which threshold should be used to decide where to stop, that is, to conclude that we have reached the correct prior and do not need to increase it anymore.

To determine the difference threshold that should be used for BIC, DIC, and PPP to inform us when to stop increasing the prior, we performed the model selection procedure presented above using different values of thresholds. Specifically, we used thresholds ranging between 1 and 80 for BIC and DIC (examining different values in this range stepwise, increasing them by increments of 1) and ranging between 0.005 and 0.3 for the PPP (examining them stepwise, increasing them by increments of 0.005). We knew which priors were appropriate (i.e., corresponding to the true cross-group differences). Therefore, we could now determine which difference thresholds values provided the most desirable results, that is, which difference threshold provided accurate information on when to stop increasing the prior.

For each threshold value, we computed the RMSE of classification and the percentage of correct classifications. The former measure (RMSE) is simply the root square error between the true prior variances defined by simulation conditions and the prior variance chosen by the testing

procedure using a particular threshold[5]. The percentage of correct classifications is the percentage of the cases where the prior was correctly identified. While this measure focused only on the exact matches, the RMSE considered also the size of the misfit from the classification, attributing a high error to large differences between the true and the chosen prior variances and a low error to small differences. We expected that a proper threshold value for the improvement of the BIC, the DIC, and the PPP fit measures would result in low values of RMSE and a high rate of correct classifications.

Outcomes

Figure 6 depicts the RMSE levels and the rates of correct classifications for different thresholds for BIC (panel a), DIC (panel b), and PPP (panel c). The RMSE values and the rates of correct classifications for different thresholds were plotted for the three scenarios (small-, middle-, and large-scale studies).

<<< Figure 6 around here >>>

Figure 6 suggests that the DIC has the potential to provide the lowest classification errors as evidenced for the RMSE values (overall, the dots presented in the left side of panel b are located at a lower position for DIC compared to BIC and PPP in the left side of panels a and c, respectively). Furthermore, panel (b), located on the right side of Figure 6, suggests that the DIC has the potential to provide also the highest percentage of correct classifications. This is followed by the performances of PPP and BIC, with the latter achieving the lowest scores of the three fit measures (see the right side of panel c and panel a, respectively).

Table 2 summarizes the specific threshold values that minimized the RMSE and the specific threshold values that maximized the percentage of correct classifications.

[5] For the classification, RMSE is defined as $\sqrt{1/R\sum_{r=1}^{R}(Prior_r - AMI)^2}$ where $Prior_r$ is selected by the procedure prior value, and AMI is the level of simulated invariance, that is, the prior. R is the number of replications.

<<< Table 2 around here >>>

In the first scenario (4x400), the suggested threshold for prior selection using the BIC fit measure was 8 according to both approaches. In other words, both approaches suggested that when searching for the correct prior variance, a BIC improvement (i.e., decrease) of 8 or higher justifies a higher prior choice, whereas a smaller change in BIC suggests not to increase the prior variance when assessing AMI. The recommended DIC threshold was 2 according to the RMSE criterion and 1 according to the rate of correct classifications criterion. The suggested PPP threshold was 0.015 using the RMSE criterion and 0.020 using the rate of correct classification criterion.

Although still very consistent with each other, slightly higher gaps between the recommended thresholds according to the two criteria are visible for the second and third scenarios of middle- and large-scale studies. For instance, the recommended threshold for DIC difference in the second (24x1,500) scenario was 14 according to the RMSE and 8 according to the rate of correct classifications criteria. In other words, the simulation suggested to rely on somewhat larger improvements in DIC for determining whether to increase the prior variance in middle-scale studies (compared to small-scale studies in scenario 1). While the recommended PPP thresholds were similar for the first and second scenarios, the recommended BIC thresholds were rather similar for the second and third scenarios. In sum, it can be concluded that for settings with a smaller number of groups and sample sizes the thresholds should be smaller whereas in larger settings one should use more strict (i.e., higher) thresholds for assessing improvement in model fit.

Finally, Figure 7 displays the percentage of selected priors (rows) in our simulated conditions for each true variance (columns) using correct thresholds for BIC, DIC, and PPP, with RMSE as a criterion for the threshold choice (results based on the percentage of correct classifications criterion are very similar and available from the first author upon request). The threshold values that were used are indicated in parentheses. The sum of the percentages in each column equals 1.0. The upper panel (a) corresponds with scenario 1 (small-scale studies), the middle panel (b) corresponds with scenario 2 (middle-scale studies), and the bottom panel (c) corresponds with scenario 3 (large-scale studies). The elements on the diagonal describe correct matches, that is, situations where chosen priors match the values that were used in the simulated data. The upper diagonal describes situations where models with too low priors were selected, and the lower diagonal describes situations where too high priors were selected compared to the true variances. For instance, in the first row of the first column in panel a, the number 0.300 indicates that the model selection procedure using BIC and a threshold of 8 chose the appropriate prior only in 30% of the cases. An ideal prior selection in each of these scenarios would be reflected in a perfect (1.0) score in the diagonal for each of the fit measures BIC, DIC, and PPP (i.e., a match between the true variance and the chosen priors).

<<< Figure 7 around here >>>

For several of the true variances conditions in all three scenarios, the results demonstrate that the correct priors are often *not* chosen when relying on the BIC fit measure. This is evident in the low numbers that appear in several locations in the diagonal in panel a in Figure 7. DIC works best, particularly for lower true variances and particularly for middle- and large-scale studies. This is evident in 100% (or nearly as high) correct prior selections in panels b and c in Figure 7 for the second and third scenarios, respectively, for true variances in the size up to

0.005, and for the condition with a true variance as large as 0.05. In the middle range (true variances ranging between 0.01 and 0.025), it is typically the DIC that does not guide us to the correct priors in the second and third scenarios. When the AMI condition equaled 0.01, DIC led us to underestimate the true prior, while when the AMI condition equaled 0.025, the DIC led us to overestimate the true variance. In the first scenario (panel a in Figure 7), the diagonal displayed the highest rates compared to the cells above and below the diagonal. In other words, in small-scale studies, the DIC was more likely to lead us to a correct rather than to an incorrect prior. However, the level of correct classifications (cells on diagonal) was not impressive and ranged between 0.696 (for an AMI level of 0.05) and 0.41 (for an AMI level of 0.025). In other words, in about 30 to 59% of the cases (depending on the AMI level), the DIC would lead us to choose a wrong prior, as evidenced in the medium to fair values on the diagonal.

Finally, the PPP performed similarly for the different scenarios in the three panels up to a true variance (AMI level) of 0.025, with middle to fair probabilities to choose the right prior. These probabilities improved with increasing sizes of studies. However, the likelihood of correct classifications (cells on the diagonal) was not impressive and ranged between 0.90 for low levels of AMI in a large-scale study and 0.30 for a level of AMI as high as 0.1 in a small-scale study.

In sum, both the PPP and DIC performed better when the size of the study increased. Given their better performance compared to the BIC, in real applications it seems advisable to use both the PPP and DIC in combination for determining the appropriate prior. According to our results presenting the slightly better performance of DIC, we can conclude that somewhat more weight ought to be attributed to this measure. Yet, for large-scale studies, the BIC could also provide correct guidance for the prior choice when the true variance is rather large (0.01 or 0.025).

Summary and Discussion

Multi-group Bayesian structural equation modeling (MG-BSEM) has gained considerable attention among substantive researchers investigating cross-group differences and methodologists exploring challenges in AMI testing. MG-BSEM modeling allows for much flexibility by applying elastic rather than strict equality constraints on item parameters across groups. This, however, requires a specification of user-defined priors for item parameters of cross-group differences. Although prior selection in general Bayesian settings is well-studied (van de Schoot et al., 2013), guidelines with respect to the selection of priors in AMI analysis are still largely missing.

The current study presented, to the best of our knowledge, the first Monte Carlo simulation to investigate the issue of tuning the normal prior variance in MG-BSEM. In two studies we examined how different prior specifications affect the results of MG-BSEM analysis and examined the difference threshold for BIC, DIC, and PPP that inform us when to stop increasing the prior variance in AMI testing in our different scenarios. It should be noted that the topic of prior selection is well-recognized in *general* Bayesian applications. Van de Schoot et al. (2013) proposed to differentiate between three types of priors: (1) non-informative priors, when no a priori information is used; (2) weakly informative priors, used for technical reasons, for example, to identify a model; and (3) informative priors with specific hyperparameters. Furthermore, the authors argued that three approaches are available for substantive researchers to formulate priors. First, one can use earlier studies and meta-analyses which contain information on cross-group differences in means, factor loadings, regression coefficients, or other parameters to define priors for those differences. Second, if no information from other studies is available, one could try to collect this information by consulting a group of experts in the field. The third

approach is the most subjective one for formulating a prior. According to this approach, the researcher may use his or her own knowledge and considerations of plausibility to formulate priors. Our treatment of priors does *not* fall into any of these categories. Instead, we used Bayesian estimation (i.e., MG-BSEM) as a practical method for relaxing equality constraints of measurement parameters in multi-group and AMI modeling. Additionally, this method helps us to find those models that reach both an acceptable model fit and recover the true latent variables' parameters of interest in cross-group comparisons. In other words, we used priors as a form of regularization (van Erp et al., 2019) of the model rather than relying on previous knowledge.

Our Monte Carlo simulation was limited to specific scenarios, as is commonly done in such studies. We explored what we considered to be three relevant scenarios in survey research: small-, middle-, and large-scale studies. First, we found that a prior misspecification had only a small influence on point estimates. This suggested that robustness checks of prior specifications based on correlating point estimates under different priors has little value for identifying the correct priors (at least using relatively small variance priors). Second, we found that well-tuned priors in MG-BSEM models resulted in more precise credibility intervals and posterior standard deviations. Thus, well-tuned priors allowed for a more accurate coverage of the mean estimation.

Next, we compared stepwise prior tuning procedures starting with low priors and testing them against models with higher priors using BIC, DIC, and PPP fit measures. We wanted to identify which change in these fit measures implies a need to further tune and increase the prior. The DIC measure proved to be overall the most effective in identifying the correct prior. In different scenarios we found that different increments of fit measures should be applied for prior tuning. It should be noted at this point that we are not claiming to have found the ultimate thresholds that could be applied in all situations for these fit measures. The coverage of possible

scenarios in the current study is too small. What we have clearly shown, however, is that the size of fit measures thresholds in AMI testing for tuning the priors is heavily dependent on the size of the study (number of groups and sample sizes). In general, the larger the research design is, the higher the threshold. Indeed, if conditions in future applied studies would be similar to those presented in the current study, the thresholds we found could be used as guidance. However, our recommended threshold criteria need to be treated with extra caution, because they apply to the specific scenarios we examined.

Indeed, while the scenarios we included covered what we consider common situations in survey research, they are not conclusive of various other realistic situations. For instance, when using survey data which include a large number of countries and time points, one would need to consider scenarios with a number of groups as high as 90 (e.g., Zercher et al., 2015) or more (e.g., when using data from the European Social Survey including 8 or more rounds and 20-30 countries). We did not include this type of scenario due to the fact that Bayesian computations are still very consuming in terms of computational power. The results presented in this paper were obtained using a high speed computational server, but the simulations required nevertheless several months to complete. Furthermore, the simulations did not cover all possible conditions also in terms of the type of bias. In the current study we examined situations in which both item loadings and intercepts were approximately invariant. However, obviously, other combinations are also possible, in which, for example, item intercepts are approximately invariant while the item slopes are fully non-invariant, or where various sets of item parameters are fully non-invariant whereas others are approximately invariant (so-called partial approximate invariance). There are other conditions of misspecifications that we did not cover in our simulations that may also be studied, such as the existence of residual correlations that, if not accounted for, may

influence the selection of priors. Thus, this study is only a small step in a direction which requires a series of additional studies covering diverse conditions and scenarios in order to be able to provide *general* guidelines for tuning the normal variance priors of the group differences in MG-CFA models in applied research. In addition, ideally, given the wide array of possible scenarios, applied researchers would perform their own Monte Carlo studies, based on the approach presented in this paper, simulating scenarios that best mimic their own research designs.

The study is not free of some other limitations. In the current study we provided recommendations on the use of thresholds for three types of fit measures: BIC, PPP, and DIC. Yet Bayesian modeling is not restricted to these three measures of fit, and other well-established indices exist, such as the leave-one-out cross-validation (LOO) (Vehtari, Gelman, & Gabry, 2017), the widely applicable information criterion (WAIC; Vehtari et al., 2017), or the data agreement criterion (DAC; Bousquet, 2008; Lek & van de Schoot, 2019). However, these fit measures are not available in commercial software packages that can run MG-BSEM models such as Mplus. The PPPP (prior-posterior predictive p-value) is implemented in Mplus, but for testing only one parameter at a time, which may be of little help for AMI analysis (see Hoijtink & van de Schoot, 2017; Asparouhov & Muthén, 2017). Obviously, in principle, MG-BSEM models could be specified also in other software packages such as STAN (Carpenter et al., 2017) or JAGS (Plummer, 2015), or even, with some limitations, in the blavaan R package (Merkle & Rosseel, 2018). However, in practical terms, it is very difficult, particularly for applied researchers, to specify BSEM models using these software packages and perform valid estimations. In fact, we are not aware of any practical application of MG-BSEM models using these packages. Nevertheless, future applications and developments of MG-BSEM invariance

modeling would profit from the use of a broader spectrum of fit measures. Moreover, we would like to emphasize that in the current study we focused on prior variance selection and assessments of model fit. However, it should be noted that a model may provide *an adequate fit with different posterior inferences under various plausible alternative models* (Gelman 2013; p. 141). Therefore, sensitivity analyses that examine changes in posterior distributions using various priors are essential. Sensitivity analysis in the Bayesian framework is well developed (e.g., Kass, Tierney, & Kadane, 1989; Oakley & O'Hagan, 2004; Weiss, 1996; Weiss & Cook, 1992) and should supplement analyses of real data.

In sum, the current study could show that an appropriate selection of priors in Bayesian AMI modeling is particularly important for estimating more precise credibility intervals and posterior standard deviations, allowing for a more accurate coverage of the mean estimation. In our conditions, particularly the DIC but also to some extent the PPP were potentially helpful for determining the right prior, whereas the BIC could also provide a correct guidance for the prior choice in larger-scale studies when the true variance was rather large.

References

Asparouhov, T., & Muthén, B. (2010). *Bayesian Analysis Using Mplus: Technical Implementation, September 29, 2010*. Retrieved from http://www.statmodel.com/download/BayesAdvantages18.pdf

Asparouhov, T., & Muthén, B. (2017). *Prior-posterior predictive p-values*. Mplus Web Notes: No. 22, Version 2. Retrieved from https://www.statmodel.com/download/PPPP.pdf

Asparouhov, T., Muthén, B., & Morin, A. J. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: Comments on Stromeyer et al. *Journal of Management, 41*(6), 1561–1577. doi:10.1177/0149206315591075

Berger, J. O., & Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics, 36*(2), 963-982. doi:10.1214/07-AOS501

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling*, *20*(3), 518-540. doi:10.1080/10705511.2013.797839

Bousquet, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics*, *35*(9), 1011–1029. doi:10.1080/02664760802192981

Braeken, J., & Blömeke, S. (2016). Comparing future teachers' beliefs across countries: Approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item functioning. *Assessment & Evaluation in Higher Education, 41*(5), 733–749. doi:10.1080/02602938.2016.1161005

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). New York: The Guilford Press

Cain, M. K., & Zhang, Z. (2018). Fit for a Bayesian: An evaluation of PPP and DIC for structural equation modeling. *Structural Equation Modeling*, 1–12.

doi:10.1080/10705511.2018.1490648

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1). doi:10.18637/jss.v076.i01

Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (2018). Testing for approximate measurement invariance of human values in the European Social Survey. *Sociological Methods & Research, 47*(4), 665–686. doi:10.1177/0049124117701478

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5*, 982. doi:10.3389/fpsyg.2014.00982

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 25–75. doi:10.1146/annurev-soc-071913-043137

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8–38. doi:10.1111/j.2044-8317.2011.02037.x

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*(4), 733–760. https://www.jstor.org/stable/24306036

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton: CRC Press.

Hoijtink, H., & van de Schoot, R. (2017). Testing small variance priors using prior-posterior

predictive p values. *Psychological Methods, 23*(3), 561–569. doi:10.1037/met0000131

Hoofs, H., van de Schoot, R., Jansen, N. W., & Kant, I. (2018). Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA. *Educational and Psychological Measurement, 78*(4), 537–568. doi:10.1177/0013164417709314

Kass, R., Tierney, L., & Kadane, J. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika, 76*(4), 663–674. doi:10.1093/biomet/76.4.663

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling, 24*(4), 524–544. doi:10.1080/10705511.2017.1304822

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). London: Academic Press.

Lee, S. Y. (2007). *Structural equation modeling: A Bayesian approach*. John Wiley & Sons.

Lek, K., & van de Schoot, R. (2019). How the choice of distance measure influences the detection of prior-data conflict. *Entropy, 21*(5), 446. https://doi.org/10.3390/e21050446

Lek, K., Oberski, D., Davidov, E., Cieciuch, J., Seddig, D., & Schmidt, P. (2019). Approximate measurement invariance. In T. P. Johnson, B. E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods. Multinational, multiregional and multi-cultural contexts* (pp. 911–929). New York: John Wiley & Sons.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543. https://doi.org/10.1007/BF02294825

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter

expansion. *Journal of Statistical Software, 85*(4), 1–30. doi:10.18637/jss.v085.i04

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074-2102. doi:10.1002/sim.8086

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313–335. doi:10.1037/a0026802

Muthén, B., & Asparouhov, T. (2013). New methods for the study of measurement invariance with many groups.Retrieved from https://www.statmodel.com/download/PolAn.pdf

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, 978. doi:10.3389/fpsyg.2014.00978

Muthén, L., & Muthén, B. (1998-2017). *Mplus User's Guide* (8th ed.). Los Angeles, CA: Muthén & Muthén

Muthén, L., & Muthén, B. (1998-2019). Mplus [computer program]. Los Angeles, CA: Muthén & Muthén.

Oakley, J., & O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society, Series B, 66*(3), 751–769. doi:10.1111/j.1467-9868.2004.05304.x

Plummer, M. (2015). JAGS version 4.0.0 user manual. Retrieved from https://sourceforge.net/projects/mcmc-jags/files/Manuals/

Pokropek, A., Davidov, E., & Schmidt, P. (2019). Assessing measurement invariance using traditional and newer approaches: A Monte Carlo simulation study. *Structural Equation Modeling*. Advance online publication. doi:10.1080/10705511.2018.1561293

Seddig, D., & Leitgoeb, H. (2018). Exact and Bayesian approximate measurement invariance. In

E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural analysis: Methods and applications* (2nd ed.) (pp. 553–579). New York: Routledge.

Sijtsma, K. (2009). On the use, misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. doi:10.1007/S11336-008-9101-0

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Steck, H., & Jaakkola, T. S. (2003). On the Dirichlet prior and Bayesian regularization. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *NIPS'02 Proceedings of the 15th International Conference on Neural Information Processing Systems* (pp. 713-720). Cambridge, MA: MIT Press.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research,* 25, 78–90. doi:10.1086/209528

Stromeyer, W. R., Miller, J. W., Sriramachandramurthy, R., & DeMartino, R. (2015). The prowess and pitfalls of Bayesian structural equation modeling: Important considerations for management research. *Journal of Management, 41*(2), 491–520. doi:10.1177/0149206314551962

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 770. doi:10.3389/fpsyg.2013.00770

van de Schoot, R., Mulder, J., Hoijtink, H., Van Aken, M. A., Semon Dubas, J., Orobio de

Castro, B., . . . Romeijn, J.-W. (2011). An introduction to Bayesian model selection for evaluating informative hypotheses. *European Journal of Developmental Psychology, 8*(6), 713–729. doi:10.1080/17405629.2011.621799

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217–239. doi:10.1037/met0000100

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods, 23*(2), 363–388. doi:10.1037/met0000162

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. doi:10.1016/j.jmp.2018.12.004

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. doi:10.1007/s11222-016-9696-4

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815-829. doi:10.1111/j.2005.0906-7590.04112.x

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society, Series B, 58*(4), 739–750. https://www.jstor.org/stable/2346111

Weiss, R., & Cook, R. (1992). A graphical case statistics for assessing posterior influence. *Biometrika, 79*(1), 51–55. doi:10.1093/biomet/79.1.51

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 209–231). New York: Guilford Press.

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: Exact vs. approximate measurement invariance. *Frontiers in Psychology, 6*, 733. doi:10.3389/fpsyg.2015.00733

Zondervan-Zwijnenburg, M., Peeters, M., Depaoli, S., & van de Schoot, R. (2017). Where do priors come from? Applying guidelines to construct informative priors in small sample research. *Research in Human Development, 14*(4), 305–320. doi:10.1080/15427609.2017.1370966

Tables and Figures

Table 1. Scenarios examined in the simulation studies.

Scenario	Number of groups	Sample size per group	Description
1) 4x400	4	400	A small cross-country field trial study or a small group comparison as part of a larger study
2) 24x1,500	24	1,500	A common cross-country study (e.g., ESS, ISSP, etc.)
3) 30x3,000	30	3,000	A large-scale cross-country study (e.g., PISA, PIRLS, PIAAC)

Note: ESS – European Social Survey; ISSP - International Social Survey Program; PISA - Programme for International Student Assessment, PIRLS - Progress in International Reading Literacy Study (PIRLS), PIAAC - Program for the International Assessment of Adult Competencies.

Table 2. Threshold values that minimized the RMSE and maximized the percentage of correct classifications.

Scenario	Value of the threshold for ΔBIC based on		Value of the threshold for ΔDIC based on		Value of the threshold for ΔPPP based on	
	RMSE	% correct classifications	RMSE	% correct classifications	RMSE	% correct classifications
4x400	8	8	2	1	.015	.020
24x1,500	20	21	14	8	.010	.025
30x3,000	17	19	32	18	.035	.045

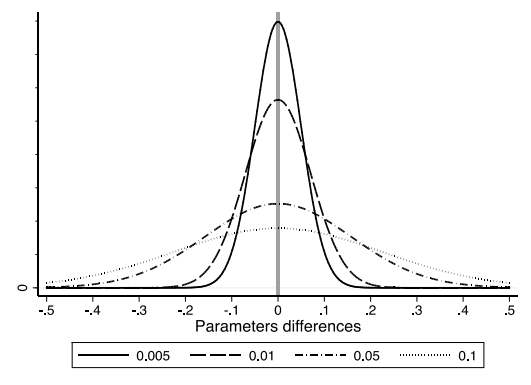

Figure 1 Expected cross-country differences of items' measurement parameters under different variance priors.

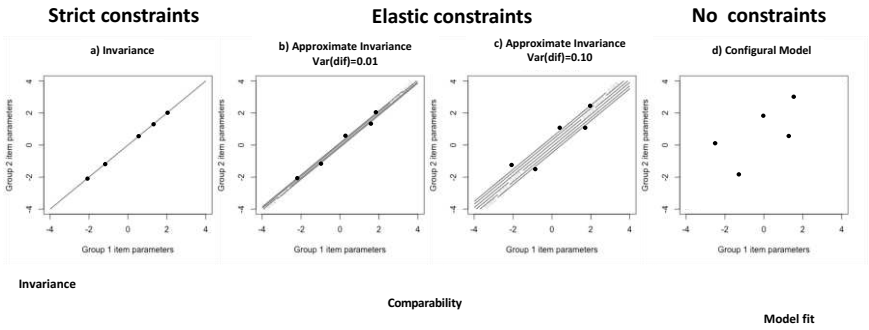

Figure 2 Expected cross-group measurement parameter differences as a function of different variance priors.

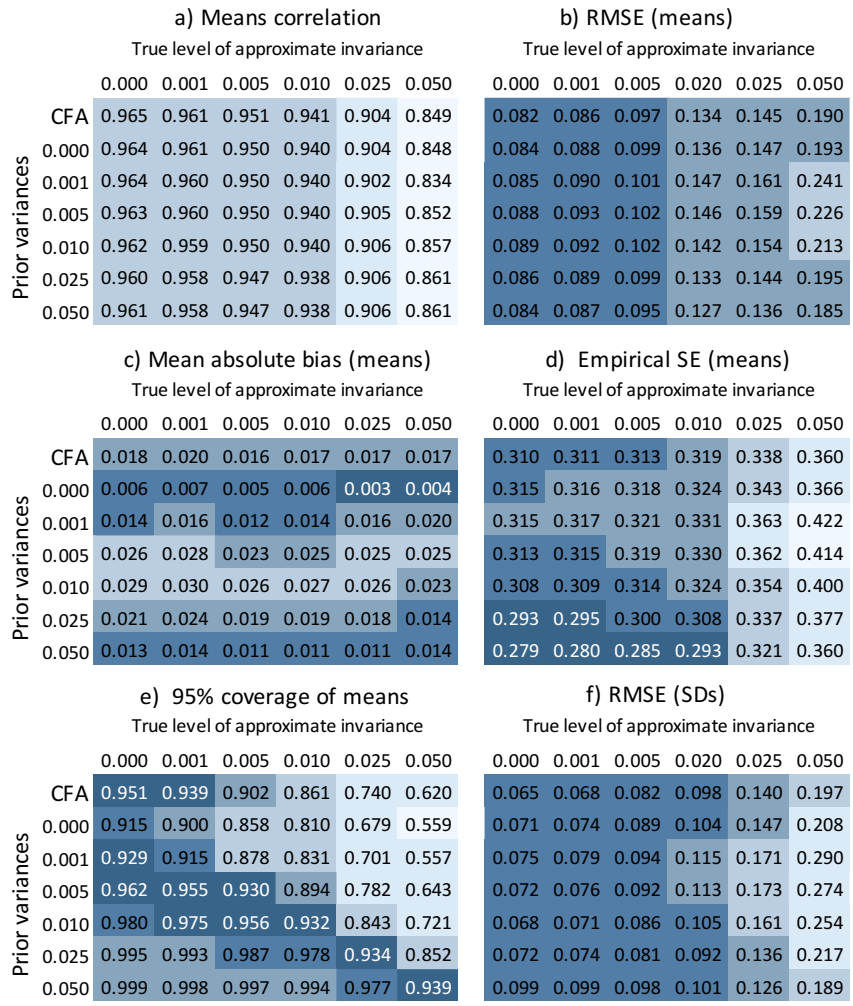

Figure 3 Consequences of different prior (mis)specifications in the first scenario.

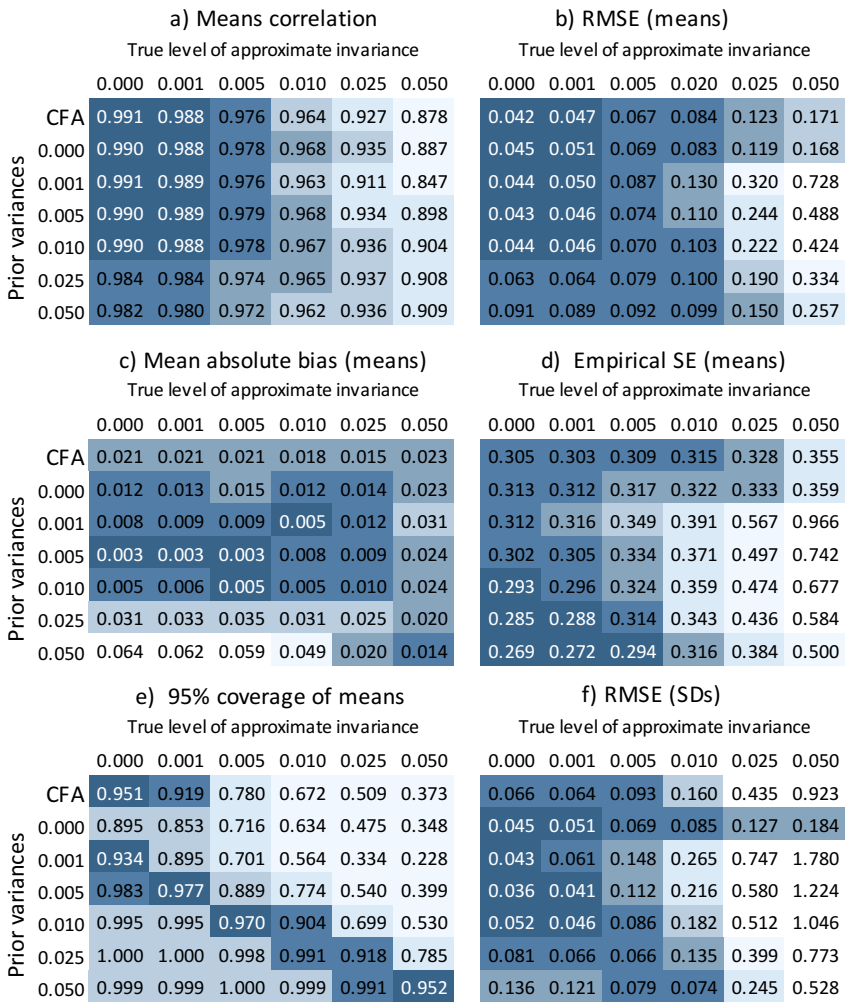

Figure 4. Consequences of different prior (mis)specifications in the second scenario.

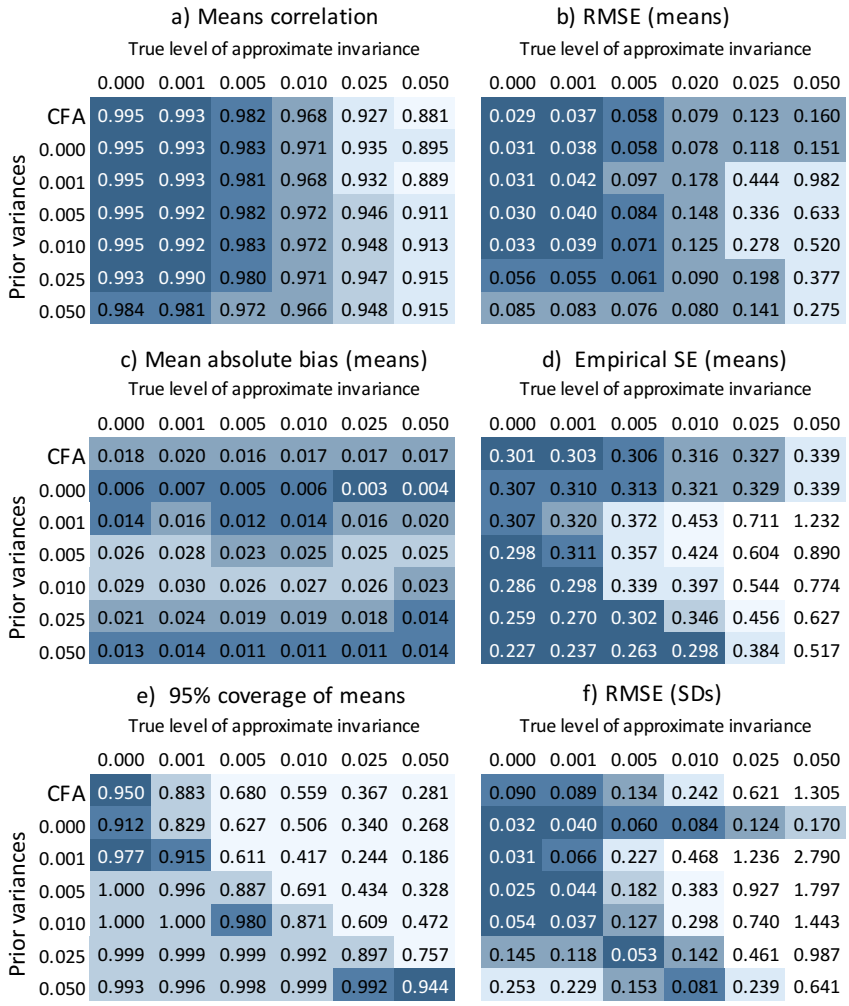

Figure 5. Consequences of different prior (mis)specifications in the third scenario.

Commented [LT1]: Do the numbers 1500 and 3000 require commas in the figure (i.e., 24x1,500 and 30x3,000)?

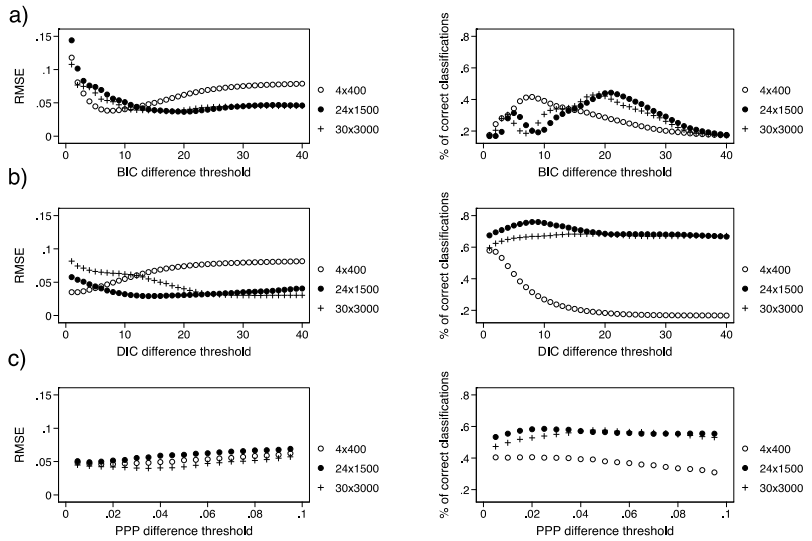

Figure 6 RMSE and the percentage of correct classifications for BIC, DIC, and PPP in three scenarios for different threshold values.

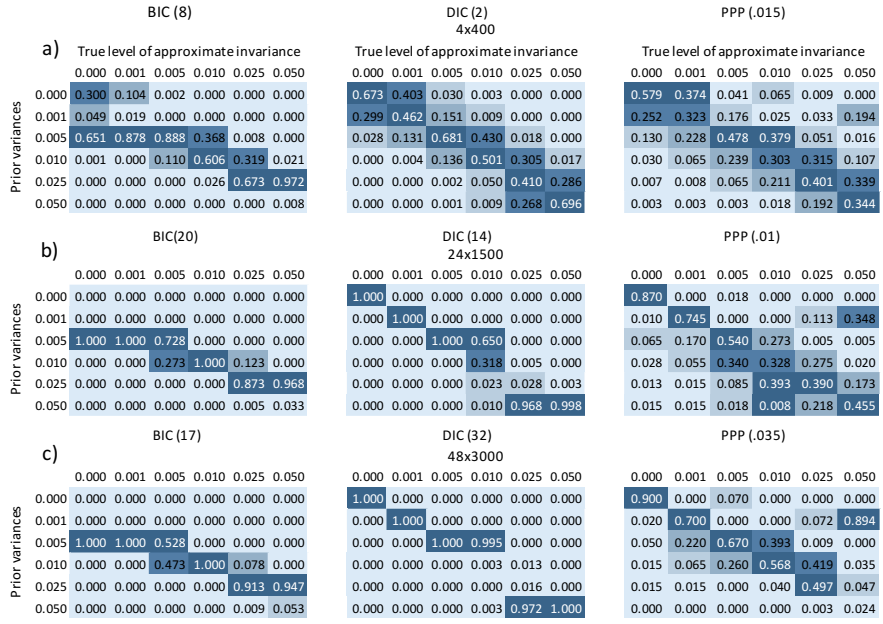

Figure 7 Percentage of selected priors for each true variance using thresholds of BIC, DIC, and PPP, with RMSE as a criterion (threshold values in parentheses). The sum of the percentages in each column is 1.0.